COMS E6998-9: Algorithmic Techniques for Massive Data — Nov 12, 2015

Lecture 19 – Sublinear algorithms for graphs

Instructor: *Alex Andoni* — Scribes: *Parthiban Loganathan*

1 MST Cost in Bounding Degree

The model we will consider is a graph G with n vertices and total degree $d = \sum_{i=1}^{n} d_i$ where d_i are the degrees of vertex v_i. We represent it with an adjacency list where each vertex points to a list of vertices and weights of the edges. Let the edge weights $\in [M]$.

We assume the graph is connected. This means the MST is well-defined and Cost of the MST $\geq n - 1$.

Theorem 1. *We can estimate MST up to a 1+ε factor in $O(M^4 d/\varepsilon^3)$ queries.*

For example, if $M = 2$, we first look at the connected components of cost 1. If we have C_1 connected components,

$$\text{MST} = \underbrace{n-1}_{\text{connecting all vertices}} + \underbrace{C_1 - 1}_{\text{connecting connected components}}$$

Fact 2. *Let C_i be the number of connected components on graph on edges of weight $\leq i$. Then,*

$MST = n - 1 + \sum_{i=1}^{M} C_i - 1$

To estimate the MST cost, we need to estimate each C_i up to δn for $\delta = \varepsilon/M$. Note the C_i are independent.

Lemma 3 (Connected Components (CC) Lemma)**.** *For any $i \in [M]$, we can design an estimator $\hat{C}_i$ such that:*

(1) $|C_i - \boldsymbol{E}[\hat{C}_i]| \leq \delta n$

(2) $var(\hat{C}_i) \leq O(\delta^2 n(C_i + \delta n))$

(3) *Number of queries* $\leq O(M^3 d/\varepsilon^3)$

MST Algorithm

(1) For $i = 1...M - 1$, estimate $\hat{C}_i$

(2) Then $\hat{MST} = n - 1 + \sum_{i=1}^{M} \hat{C}_i - 1$

Now let us prove the above theorem.

Proof.

(1)

$$
\begin{aligned}
|\mathbf{E}[\hat{MST}] - MST| &\leq \sum_{i=1}^{n} |C_i - \mathbf{E}[\hat{C}_i]| \\
&\leq \delta n M \\
&= \big(\frac{\varepsilon}{M}\big) n M && \text{(since } \delta = \varepsilon/M\text{)} \\
&= \varepsilon n
\end{aligned}
$$

(2)

$$
\begin{aligned}
\mathrm{var}(\hat{MST}) &\leq \sum_{i=1}^{M-1} \mathrm{var}(\hat{C}_i) \\
&\leq O(\delta^2 n\big(\sum_{i=1}^{M} C_i + \delta n\big)) \\
&\leq O(\delta^2 n(nM)) \\
&= O(\varepsilon^2 n^2)
\end{aligned}
$$

Finally apply Chebhyshev's Inequality to obtain a bound with constant probability. □

Proof. CC Lemma Proof:

Input is a graph H ($= G$ on edges with cost $\leq i$). Output is $C = C_i$ = number of connected components in H. Our goal is to obtain an estimator $\hat{C} \approx C$.

Define vertex v such that $\alpha_v = \frac{1}{\text{size of CC of v}}$

We sample v in order to estimate α_v. To compute α_v, we need to find size of CC of v which may be the entire graph making the algorithm linear. So instead, we estimate α_v by thresholding it.

$\hat{\alpha_v} = max\{\alpha_v, \delta\}$

$|\sum_v \hat{\alpha_v} - \sum_v \alpha_v| \leq \delta n$

Algorithm to compute each $\hat{C}_i$

(1) For $i = 1...k$ where $k = 1/\delta^2$ pick random v_i

(2) Compute $\hat{\alpha_{v_i}}$ via Breadth First Search stopping after we see $1/\delta$ vertices

(3) $\hat{C} = \frac{n}{k} \sum_{i=1}^{k} \hat{\alpha_{v_i}}$

(1)

$$\begin{aligned}\mathbf{E}[\hat{C}] &= \frac{n}{k}\sum_{i=1}^{k}\mathbf{E}[\hat{\alpha_{v_i}}] \\ &= \sum_{i=1}^{n}\hat{\alpha_{v_i}}\end{aligned}$$

(2)

$$\begin{aligned}\mathrm{var}(\hat{C}) &= \frac{n}{k}\sum_{i=1}^{n}\hat{\alpha_{v_i}} \\ &= \frac{n}{k}\sum_{i=1}^{n}\hat{\alpha_{v_i}} \\ &= \frac{n}{k}(\underbrace{C}_{\sum\alpha_v} + \underbrace{\qquad\qquad\qquad\qquad}_{\text{max difference between } \sum\alpha_v \text{ and } \sum\hat{\alpha_v}}) \\ &= \frac{n}{k}(C+\delta n)\end{aligned}$$

(3) Number of queries $\leq k \times depth \times d = \frac{1}{\delta^2}\frac{1}{\delta}d = dM^3/\varepsilon^3$

□

The best known bound is $O(dM\varepsilon^{-3}\log\frac{dM}{\varepsilon})$ [Chazelle-Rubinfeld-Trevisen].

2 Estimating Average Degree

Problem statement:

(1) $m = n\bar{d}$ where $\bar{d}$ is the average degree

(2) Degrees are unbounded

(3) $\bar{d} \geq 1$ (ie. at least n edges in G)

The trivial solution uses $O(n)$ queries by simply iterating over all vertices and computing the sum of degrees in order to find the average. To do better, we will attempt to sample some subset of vertices in order to estimate $\bar{d}$.

First, we see that we can't compute $\bar{d}$ with constant number of queries. For example, consider a case where we do not sample a very "heavy" vertex with high degree that contributes a lot to $\bar{d}$. Or consider the case where we have $\sqrt{n}$ connected vertices and $n - \sqrt{n}$ unconnected ones. The query complexity is $\Omega(\sqrt{n})$.

Theorem 4. *We can estimate average degree $\bar{d}$ up to a $1+\varepsilon$ factor in $O(\sqrt{n}/\varepsilon^2)$ queries.*

Algorithm to compute $\bar{d}$

(1) Sample edges $e_1, ..., e_k$ iid from distribution $\{p_e\}$

(2) Estimator $\hat{d} = \frac{1}{k}\sum \frac{1}{np_e}$

We sample $\{p_e\}$ as follows:

(1) Sample random vertex u and then sample a random neighbor v along edge $e = (u, v)$.

(2) Estimator $\hat{d} = \frac{1}{k}\sum \frac{1}{np_e}$

Let u and v have degree d_u and d_v respectively. Probability of sampling a vertex u is $1/n$. Probability of then sampling a neighbor is $1/d_u$. Hence $p_e = \frac{1}{nd_u} + \frac{1}{nd_v}$.

$$\begin{aligned} p_e &= \frac{1}{nd_u} + \frac{1}{nd_v} \\ &\geq \frac{1}{n}\max\{1/d_u, 1/d_v\} \\ \implies \frac{1}{p_e} &\leq n\min\{d_u, d_v\} \end{aligned}$$

Need to show:

(1) $\mathbf{E}[\hat{d}] = \bar{d}$

(2) $\mathrm{var}(\hat{d}) = \frac{1}{k}\mathrm{var}(\frac{1}{np_e})$

$$\begin{aligned} var\Big(\frac{1}{np_e}\Big) &\leq \mathbf{E}\Big[\Big(\frac{1}{np_e}\Big)^2\Big] \\ &= \frac{1}{n^2}\sum_e \frac{p_e}{p_e^2} \\ &= \frac{1}{n^2}\sum_e \frac{1}{p_e} \\ &\leq \frac{1}{n^2} n \sum_{e=(u,v)} \min\{d_u, d_v\} \end{aligned}$$

Attempt 1: This does not work.

$$
\begin{aligned}
var\Big(\frac{1}{np_e}\Big) &\leq \frac{1}{n}\sum_{e=(u,v)} \min\{d_u, d_v\} \\
&\leq \frac{1}{n}\sum_{u} d_u^2 \\
&\leq \frac{1}{n}\big(\frac{m}{n}n^2\big) && \text{(in the case where each vertex has degree } m\text{)} \\
&= m
\end{aligned}
$$

$\text{var}(\hat{d}) = \frac{1}{k}\text{var}(\frac{1}{np_e}) = \frac{m}{k}$

$\implies \hat{d} = \bar{d} \pm \sqrt{\frac{m}{k}}$

We want $\text{var}(\hat{d}) \leq \varepsilon\bar{d}$

$$
\begin{aligned}
&\sqrt{\frac{m}{k}} \leq \varepsilon\bar{d} \\
\implies\ & k \geq \frac{n\bar{d}}{\varepsilon^2\bar{d}^2} \\
&= \frac{1}{\varepsilon^2}\frac{n}{\bar{d}}
\end{aligned}
$$

If $\bar{d} = 10$ for example, $k \approx n$ and it's linear. Hence this attempt fails.

Attempt 2: There are at most m/M vertices with degree $\geq M$. Let us call them heavy nodes.

$$
\begin{aligned}
\operatorname{var}\Big(\frac{1}{np_e}\Big) &\le \frac{1}{n}\sum_{e=(u,v)} \min\{d_u, d_v\} \\
&= \frac{1}{n}\sum_u \sum_v \min\{d_u, d_v\} \\
&= \frac{1}{n} \sum_{\substack{u \\ \text{either } u \text{ or } v \text{ not heavy}}} \sum_{v} M + \frac{1}{n}\sum_{\substack{u \\ v \text{ is heavy}}} \sum_v d_u \\
&\le \frac{1}{n} mM + \frac{1}{n}\sum_u d_u \frac{m}{M} \\
&\le \frac{m}{n}(M + \frac{m}{M}) \\
&\le \frac{m}{n}\sqrt{m}
\end{aligned}
$$

If degree was constant $m = n$, $\operatorname{var}\Big(\frac{1}{np_e}\Big) \le \sqrt{m}$.

In general, $\hat{d} = \bar{d} \pm \sqrt{\frac{m^{3/2}}{nk}} = \bar{d} \pm \varepsilon \bar{d}$ for $k = \frac{n}{\sqrt{m}\varepsilon^2}$.